WHITE PAPER

云架构设计简介

基础设施即服务（Iaas）的四项首要设计原则

什么是云

云计算是一种汇集多种技术和实践的集合，用于概括计算机硬件的配置与管理。其目的是简化终端用户的体验，令其按需获取计算机资源，用云计算的话讲，就是“一种服务”。组成一个云的资源，一般是通过符合三种抽象层次之一的某一接口提供给终端用户的。这三种抽象层次按从最具体到最抽象分别为：基础设施即服务Infrastructure-as-a-Service (Iaas)，平台即服务Platform-as-a-Service (PaaS)，和软件即服务Software-as-a-Service (SaaS)。

应用的存储、管理和更新几乎完全进行在云上，只是通过网络浏览器或瘦客户端提供终端用户服务的，被称为以软件为服务（SaaS）。采用SaaS模式部署的应用，一般是社交，合作，媒体和内容管理等应用类型。不过，我们也看到越来越多的传统桌面应用也正转向这种模式。实际的例子范围很广，比如CRM的有salesforce.com ,email如gmail,以及基于互联网的游戏的，像FarmVille 等等。

以平台为服务(PaaS)的抽象程度处于SaaS之下。所谓平台是一个为加速应用的开发而量身定做的环境。该平台包含为容纳应用的某些特性以及一组软件，如web 服务器、数据库、负载均衡等所必需的计算能力的一个规范。平台有了这些预编译的组件，使得部署和管理这些资源变得简单，因为可以把部署和管理这些资源的责任从应用开发人员那转移到云维护者（cloud maintainer）上。对一些公共平台元素，例如apache web-servers, SQL 数据库，负载均衡，平台则提供了现成的部署。

以基础结构为服务(IaaS) 则创建虚拟的硬件资源，包括虚拟机，虚拟网络和虚拟存储。IaaS与虚拟概念紧密相关，通常是更高一级抽象层次如PaaS和SaaS的基础。尽管存储在逻辑上是IaaS的一部分，它却经常被单独考虑。越来越被广泛采用的,没有其他IaaS部分而只有云存储的情况则尤其如此。这样的例子有Amazon S3 和 Rackspace Cloud Files ,以及 SaaS 应用像dropbox 和 box.com.

本文概述了部署IaaS 云时四个主要设计考虑，并且考察了采用Mellanox的以太网和InfiniBand interconnects建立IaaS的好处。

为什么要选择Mellanox组云？

Mellanox Technologies是RDMA技术世界领导者。RDMA是一种网络适配器功能，允许通过网络连接到一起的商用x86系统间进行内存到内存的传输。在所有市面上的互联技术中，Mellanox的 RDMA允许以最高的带宽，最低的延迟和最小的CPU周期进行虚拟机之间的网络和存储数据传输。Mellanox的互联技术主要为需要建立可扩展,低成本的IaaS的云提供商提供便利。本文主要描写对IaaS的优势，但是，由于SaaS和PaaS是建立在IaaS基础之上，所以，对这两个层次的好处也一脉相承。

建立IaaS云的设计考虑

通过与众多客户的深度交流，数据中心的建立，以及与IaaS构架师和管理人员紧密合作，Mellanox发现全世界最好的云数据中心所具有的几点共同特点。这些共性分别在如下所示的四个设计原则中进行阐述。

原则一：可扩展的物理设计

最好的IaaS云具有快捷简单的部署物理机器的方式，并能立即将其作为虚拟基础结构的一部分开始使用它们。许多公有和私有云一般在初始时规模较小，随着越来越多用户开始将应用放在云上，而需要增加额外的容量。

采用一个高密度节点是首要目标。 第二个目标是高度模块化设计。这两个目标相互关联，因为它们允许新资源以渐进的方式加到云上，同时确保配置简单。通常高密度节点在热和电方面效果更佳，回报率更高。一般来说，相比于多机架单元的同类产品，高密度节点的采购价格差不多，有时还更便宜。这些模式的存储能力和IO扩展性，是IaaS架构师无法部署高密度方案的两个主要限制。

在向现有的云上部署新的物理系统时，避免造成现有系统停机是一个目标，由于新系统安装的复杂性，这种情况时有发生。当设计一个云集群时，一个重要的考虑是，新节点安装后，要避免由于重新执行负载均衡，而导致整体性能下降。

从组网的角度，当然要优选可以同时使用密集型（dense form factors）并且能降低物理安装难度的技术和拓扑结构。采用高带宽的互联方案，可以在最少的物理连接上部署最大数量的虚拟基础结构设备。另外，这将不同网络流量类型（如，管理，存储，网络）聚合到一根单独的网线上，简化了系统安装。

假设某个给定的虚拟基础结构设备需要大约1Gb/s的连接能力，包括网络，存储和管理流量，这大概还是保守的估计。如果该IaaS是设计从虚拟机主机上分离存储资源（原则3）的话，这个数字就更加保守了。根据这个假设，大约需要n Gb/s的带宽，这里n是运行在任意给定VM主机上运行的虚拟机的数量。除非该IaaS是设计用来运行科学和处理用Nehalem或之后组件的计算作业类型的。在一般情况下，一个典型的云部署将会在任意地点，在每个物理基础结构服务器运行从15到30个左右的虚拟机(VMs)。在单一端口上支持该量级带宽的唯一技术是40GbE以太网或40-56Gb/s InfiniBand。

具有在单一端口上提供这种需求能力的方案非常重要，因为密集型系统设计有这样的要求。正如之前所讨论的，该方案在资本和运营费用上，都具有优势。此外，由于不需要再为聚合和冗余绑定链路，它也简化了配置。

这种密集组网类型的一个理想形式是，只需把网卡（NIC）直接做在主板上（LOM） 或做成一个特定的子卡(如Mezz 或ALOM)就可以了。实际的例子包括具有主板40 Gigabit InfiniBand的HP SL 390，或具选配Mellanox 40Gb/s 以太网或InfiniBand子卡的Dell C6100。在这两个例子中，除了密集型网卡，系统中的PCIe插槽依旧保持开放，允许扩展。

原则二： 简单的配置规则

最好的IaaS云以快速简单的方式部署新的虚拟机，并把他们连接到其余的虚拟基础结构上。这项任务包括识别正确的管理程序hypervisor以创建虚拟机，为虚拟机提供所需的存储流量以及之后提供虚拟机（VM）所需的网络连接能力。调度是把虚拟基础结构映射到物理资源的行为，也是云设计中最具挑战的。简单的配置方法意味着调度选项是有限的，无关紧要的或是自动的。

自动还是手动在云上执行配置操作，很大程度上取决于部署的规模，用户数量以及分配上动态还是静态的程度。最好的解决方案，是在任意模式下可通过高度集成的中央管理系统进行工作。

Mellanox的 UFM软件提供了此类型的解决方案。UFM给管理员提供了针对节点所有方面的控制，包括配置，QoS和vNIC 管理， 以及通过网络远程管理固件和软件版本的能力。UFM 也为一般任务提供了简化的自动操作。当网络发生改变时，比如VM 迁移，UFM将自动地配置物理和虚拟网络单元，以保证连接和策略（policy）的完整性。UFM在兼顾到网络调度的同时，也集成了一组工业标准化的自动调度器，从MOAB到OpenStack，以处理非网络调度。

今天，大多数IaaS方案都拥有对组件，如CPU和内存，保证服务级别协议（SLAs）的能力。然而，只有最好的IaaS部署才考虑和规划为云内的组网提供服务级别协议。UFM有独特的性能监控引擎，可以监控性能流量的潜在问题，如流量瓶颈等方面提供近实时的信息。这些信息会自动关联到UFM代表云应用和服务的逻辑模型上，用户最终能够在一个地方看到，放在网络上的云服务到底是什么样的服务等级，它们所占用的带宽，以及存在的任何流量问题。 再与微调功能（tweaking） 和配置能力一起，一个完整的操作周期就完成了。

InfiniBand 技术也为云提供商提供了一些关键优势。InfiniBand是一个自愈型网络连接（fabric），即其中央管理系统不断地，自动监测网络链路错误和拥塞链路。该技术通过基于行业标准的方法，根据实时信息来动态平衡和纠正网络路由。另外，InfiniBand在规模上具有优势，相比以太网一个子网几百个节点，其一个子网上可以增加至多达上千个节点。这简化了网络，同时降低了配置路由器时的复杂度和减少了错误。最后，InfiniBand提供了无损耗的网络，这对在网络协议上支持存储，和避免网络拥塞都非常重要。

原则三： 弹性设计

一个IaaS系统应该具备适应负载需求不断变化的能力。该IaaS应该具备随着负载增加扩大规模，随着负载降低压缩规模的能力。工作负载也同样会产生动态的变化，能够做负载再均衡也至关重要。从IaaS的角度，这意味着能够重新调整虚拟机的物理位置，或者调整存储区块的位置，来实现存储子系统之间平衡负载。

云基础结构层经理所能够做出的一个最重要的设计决策，就是区分开运行VMs的节点与存储节点。这会产生三大优势：首先最重要的，这将允许有效的分配存储。当存储被聚集时，就不会发生碎片。其次，运行VM镜像的调度再也不用局限于节点的本地存储子系统了。这使得分配变得简单。最后，这也允许最大的弹性，因为只需要对运行的镜像在线迁移，而不需要包括存储区块。

提供分离存储的挑战在于网络性能。运行远程文件系统或基于块存储的协议，如NFS或

ISCSI，对每个VM都需要额外的带宽。然而，仅仅带宽还不够，因为存储的远程访问的时间也会有影响，除非采用一个低延迟的网络。对于这类任务，分布式文件系统也很受欢迎，因为该文件系统能把这种带宽需求分摊到多个节点上。然而，这种额外的通信将会增加跨网络的额外负载。为了减轻这种影响，又不产生一个全然独立的存储网络，你需要设计采用一个具有强大的流量分离机制和QoS规则（rules）的网络。

在区分VM基础结构节点与存储节点时，最后一个风险是处理存储流量时会增加CPU的利用率。采用为存储流量提供了卸载能力（Offload），甚至RDMA技术的网络适配器，解决了这个问题。

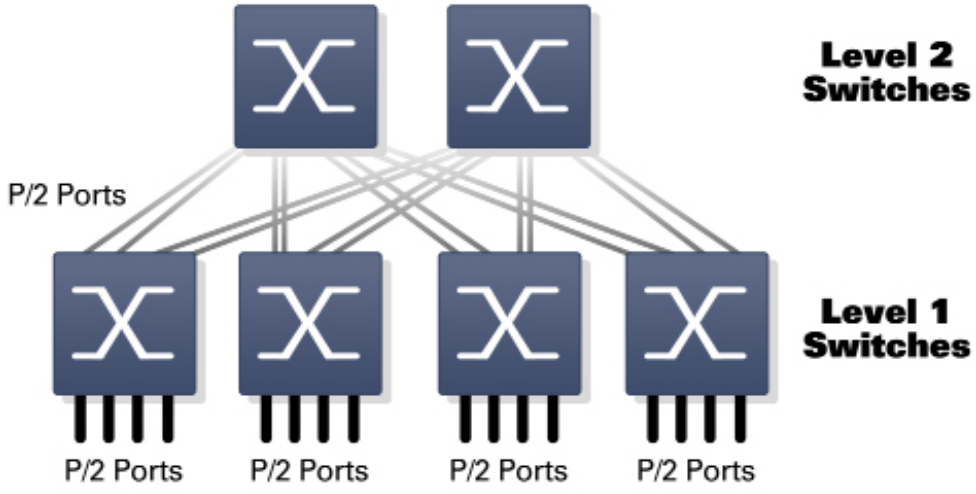
一旦存储从实时迁移过程中解脱出来，该云便具有了在基础结构节点上更加灵活地重新分布负载的能力。最终剩下的挑战就是如何把迁移时间尽量缩到最短，考虑到这是个需要高CPU，高网络带宽的操作。做这个工作时，如果拥有一个低延迟，高带宽的互联方案，例如40Gb/s 以太网或56Gb/s InfiniBand，将使得VM在线迁移的时间比传统的万兆以太网（10GbE）网络最高快至4倍。(Beck, 2011)

原则四：支持东西向网络流量的能力

南北向流量是从管理程序hypervisor，通过架顶式交换机（Top of Rack）然后经过汇聚交换机，流出到互联网上的流量。该流量一般是终端用户或WAN上产生的。然而研究表明，云上的主要的流量是东西向，即在云内部的虚拟服务器之间的流量。这很大程度上是由于，一个SaaS应用的标准构成，将运用许多虚拟结构和平台实体，以建立一个能在期望的响应时间内提供动态内容的应用。平台的一些元素如内存缓存服务器，负载均衡器以及数据必须在基础结构实体，例如存储单元或云文件，搜索数据，处理和汇聚这些信息（Morgan, 2011）。所有这些处理都发生在云的内部，将会产生相比于最终为了响应WAN的北部边界庞大很多的东西向流量。

当云的基础结构失衡，资源碎片遍布物理基础结构时，往往在东西流量方面会做出“错误”的调度决策。在事先不知道哪个虚拟基础结构设备需要交互时，唯一能够确保东西向流量流工作正常的方法就是采用一种能确保恒定的跨区带宽和延迟的网络拓扑，比如“胖树”结构。

胖树Fat-Tree 结构区别于传统的3级数据中心结构的方面在于，它的架顶层（Top of Rack）与集成层是以全互联的方式设计的。该树若不是完全无阻塞，就是阻塞率很低。对于复合型（hyperscale）的网络部署，一个汇聚的胖树（FAT-Tree） 可以绑定在POD内，并且集成成一个第三级（a third tier）。然而，在这种配置下，客户端应需要在一个单独的POD上绑定，以确保他们能够利用到胖树上.



摘要

通过在IaaS设计中引用以上四个原则，就能部署一个高度优化并有效的云在设计一个IaaS。通过运行在一个高度动态的IaaS环境，这种方法确保了最好的硬件使用率。Mellanox的硬件和软件产品完美支持这些原则，提供了最好的投资回报率和总体拥有成本。

原则 #1 – 可扩展的系统设计

- 考虑高密度和模块化设计，在可选的网络方案中选择主板集成或者Mezz卡
- 将网络，存储和管理IO汇聚到单一网络

原则 #2 – 简单的配置规则

- 采用同时支持自动或手动配置的工具
- 考虑采用提供动态路由和自愈能力的InfiniBand

原则 #3 – 弹性的设计

- 分离基础结构节点与存储节点
- 采用高带宽的网络互连方案来提高实时迁移速度4倍

原则 #4 – 支持东西向网络流量能力的规划

- 采用胖树（Fat-Tree）拓扑结构，避免东西向流量拥塞
- 选用InfiniBand在单2层网络可以搭建多达上千个节点

特性	Mellanox FDR (56Gb/s) InfiniBand	Mellanox 万兆/4万兆	传统千兆/万兆网
低延时	最低	低- VMA或RoCE	高
无损操作模式	直接支持	通过DCB支持	通过DCB支持
中央管理	是	是	否
支持自动调度器	是	是	可能
支持RDMA存储	是	是	可能
硬件增强碎片分离	是	是	否
自愈能力	是	否	否
动态负载均衡	是	否	否

引用

Beck, M. (2011). VM Migration Acceleration over 40GbE.

Retrieved from http://www.mellanox.com/pdf/PPT/VM%20Migration%20over%20 40GigE.pdf

Morgan, T. P. (2011, 11 29). The Register. Retrieved from The Register:

w://www.theregister.co.uk/2011/11/29/cisco_cloud_data_center_traffic_index/

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com